History of Clinical Prediction (Exhibit 2)

Up until the 1980's, the predominant tool in use for making predictions of future sex offenses (or other forms of violent, dangerous behavior) was clinical judgment. Though there was a mounting body of research that challenged this approach, there were no well established, generally accepted reasons to question this methodology; it certainly seemed reasonable to turn to experts who had treated and studied sex offenders when it came time for making release decisions. Surely, they were more likely to know whether an offender was dangerous than someone whose first contact with an offender was as a juror. After years of experience—sometimes with hundreds of sex offenders—and careful study of the literature that built upon their years of formal training, such experts were also expected to have more pertinent wisdom regarding this matter than a judge who presided over a relatively small number of such cases.

This was still the prevailing view when—despite the testimony that was presented to challenge the prevailing wisdom—the Supreme Court issued its *Barefoot* decision. The challenge failed to convince the SC of the unscientific and invalid nature of the predictions it was accepting even though, years earlier, researchers had begun to show that assumptions of such clinical expertise were usually *invalid*. The assumed expertise was *almost always* found to be an illusion whenever it was subjected to close examination. For example, in an early classic study examining clinical expertise, Goldberg (1959) asked novices and highly experienced, Ph.D. psychologists to look at MMPI[1] scores from individuals tested prior to their discharge from a mental hospital. The clinicians were asked to use the MMPI results to determine the severity of the discharge diagnosis (neurotic vs. psychotic, a much simpler distinction than will vs. won't commit a new sexual offense). Goldberg had previously constructed an actuarial rule based on his studies of the MMPI profiles of patients whose discharge diagnoses he already knew. The formula—which since has come to be known as "Goldberg's Rule" (Dawes, Faust, and Meehl, 1989)—was to add up the scores on three MMPI scales and subtract the scores on two others. According to Goldberg's (simple) Rule, if the result was 45 or more, it was predicted that the individual would be discharged with a psychotic diagnosis. The vast amount of information available in the rest of the MMPI and the complex patterns formed by the many scales were ignored in this simple actuarial measure.

The result? The novices and the experts using "clinical judgment" performed similarly averaging 62% correct diagnoses:

[1] The MMPI is *the* most researched and studied psychological test; there have been thousands of carefully controlled studies using the MMPI and it has been administered to many hundreds of thousands of people. It consists of approximately 560 true-false questions about a wide variety of behaviors, feelings, beliefs, values, experiences, etc. The pattern of answers an individual gives is then compared to see if it matches patterns obtained when the test was given to carefully selected groups, e.g., paranoid schizophrenics, severely depressed people, individuals with obsessive-compulsive disorder, etc. Scales have been developed that measure syndromes or character patterns like "depression"; for example, if a person answered "true" to a question like "I often wake up dreading the day ahead," he would get one point for a depression scale. If the answer was "false" to something like "Every now and then I have a really good day," one point would be scored on the depression scale. 20 or 30 items would be selected for such a scale based on the pattern of actual responses that had been given by people who were known to be suffering from depression. The higher the score on the depression scale, the more depression it indicates. Hundreds of scales have been developed including paranoia, hypochondriasis, phobias, social conformity, etc. Patterns of high scores on the different scales scored are then used to make more specific conclusions about individuals.

> One surprising finding—that amount of professional training and experience of the judge does not relate to judgmental accuracy—has appeared in a number of studies.[2] (Goldberg, 1959)

The single most accurate judge made 67% correct decisions. The Goldberg Rule achieved an accuracy rate of 70%. Goldberg then trained the judges by giving them 300 MMPI profiles to practice with and providing immediate feedback on accuracy. Even after 4,000 practice judments, *none* of the judges equaled the Goldberg Rule. Finally, Goldberg gave some of the judges (including all of the experts) the information from the Goldberg Rule on each case and allowed them to use the extra information (that the rule in its simplicity could not make use of) and their clinical judgment to use or modify the prediction made by the rule. Though there were some gains in accuracy, no judge did as well as the rule: Every judge would have been more accurate if they avoided using clinical judgment and always used the rule alone.

Doubt about clinical judgment built upon the work of Meehl (1954) who had started the wave of controversy when he first questioned the expertise of experts in predicting human behavior. Meehl's book served the same function as the child who dared to question the emperor's new "clothing." Slowly, the research questioning the clinical expert's, legal clothes picked up momentum so that, by the 1970's the tide among *scholars* (i.e., scientists and researchers who were not practitioners who made a substantial part of their livelihood from making clinical predictions) was beginning to turn.

> Many clinicians have been making unreliable and invalid judgments based on invalid premises, illogical assumptions, unproven relationships, inappropriate applications of unproven theories and other types of error. (Thorne, 1972)

By the late 70's, the waves of mounting evidence began to take effect resulting in the "Report of the Task Force on the Role of Psychology in the Criminal Justice System."

> [T]he validity of psychological predictions of violent behavior [is] . . . so poor that . . . one could [say] . . . that psychologists are not professionally competent to make such judgments.

But it wasn't until the early 80's that the tide began to turn decisively. Monahan (1981), in an influential work, reported that clinicians were wrong in two out of three predictions of violence. The *amicus* brief filed by the American Psychiatric Association in 1983 (*Barefoot*) was based, in part, upon Monahan's work. This is where the APA concluded that "psychiatric predictions of long-term future dangerousness are wrong in two out of every three cases."

By the mid 1980's, the new understanding had begun to reach general acceptance among the scholars and professional associations. At this time, Meehl (the fellow who started all the trouble) could look back and, comparing actuarial versus clinical prediction, conclude:

[2] And it has since been confirmed in many studies (Dawes, et al., 1989; Goldberg, 1959; Menzies, Webster, & Sepejak, 1985; Monahan, 1981; Faust & Ziskin, 1988; Mossman, 1994; Menzies, Webster, McMain, Staley, & Scaglione, 1994; Rice & Harris, 1995).

> There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one. (1986, pp. 373-374)

There is a tendency now to talk of Pre-Monahan (1981)—again, Monahan was the researcher that established that clinical predictions of dangerousness are wrong 2 out of 3 times and that clinicians are often poorer predictors than lay people—and Post-Monahan research. By the end of the 1980's, Dawes, et al. (1989) captured the prevailing view that had finally been established—general acceptance had become clear—among scientists in the field. They reviewed and summarized the studies comparing clinical and actuarial judgment. They concluded that there were "nearly 100 comparative studies in the social sciences. In virtually every one of these studies, the actuarial method has equaled or surpassed the clinical method" (p. 1669). By the early 1990's, this was no longer controversial: All of the experts who study methods of predicting this type of human behavior have concluded that actuarial/statistical prediction is more accurate than clinical prediction. This is so well established, I have been unable to unearth a single violence prediction study carried out in the last decade that even bothers to study or use the old-style clinical prediction. The standard goal of research has now become improving statistical prediction.

Many of the professionals *who make their living* from testifying about future dangerousness do make use of clinical prediction of the type that has been discredited (see Thorne, 1972; Epperson, Kaul, and Hesselton, 1995: Rice & Harris, 1995; Borum, 1996; Rice, 1997), though they now tend to cloak this clinical prediction under the rubric of the "guided clinical" or even the "adjusted actuarial" method. This should not be taken to mean that there has been *any* acceptance in the general psychiatric or psychological community of the validity of such clinical methods. In fact, the reverse is true. Clinicians who engage in this work comprise a tiny fraction of practicing clinicians. And their professional associations have not changed their positions from the highly skeptical and ethics questioning *amicus* brief (1983) submitted by the American Psychiatric Association in *Barefoot* and the 1978 "Report of the Task Force on the Role of Psychology in the Criminal Justice System" published in the American Psychological Association's primary organizational journal (*American Psychologist*, 1978), which stated:

> the validity of psychological predictions of violent behavior, at least in sentencing and release situations we are considering, is extremely poor, so poor that one could oppose their use on the strictly empirical grounds that psychologists are not professionally competent to make such judgments.

More recently, one finds these positions affirmed again and again in the same organs of these associations. For example, in the *American Psychologist* we find:

> Assessments of dangerousness made by clinicians continue to ignore the research on the prediction of violence. (Borum, 1996)

> Most assessments of dangerousness were (and still are) based exclusively on unaided clinical judgment. In almost every situation in which they have been studied, actuarial predictions have outperformed unaided human judgement

> (Meehl, 1954, 1986, 1996). . . . Mossman (1994) recently showed again the superioritv of actuarial methods over clinical methods for long-term predictions in a meta-analysis of studies of predictions of violence specifically." (Rice, 1997, p. 416)

The American Psychiatric Association Task Force Report On Sexually Dangerous Offenders (Zonana, et al,, 1998) and a Brief of *Amicus Curiae* American Psychological Association (DeBruin, 2005) also demonstrate a fundamental lack of change between their more current professional associations' stance and the psychiatric association's 1983 *amicus* brief and the psychological association's 1978 task force report.

Furthermore, just as Meehl noted the unequivocal uniformity of the empirical results, there is virtual unanimity among those who study the accuracy of behavioral prediction, in general, and the prediction of violence and/or sexual offenses, in specific. Actuarial/statistical prediction beats clinical prediction every time.

> Indeed, a large body of literature over the last several decades has consistently demonstrated the general superiority of actuarial prediction over clinical prediction in virtually every decision-making situation for which the issue has been studied. (Harris, Rice, & Quinsey, 1993)
>
> In virtually every decision-making situation for which the issue has been studied, it has been found that statistically developed prediction devices outperform human judgments. (Gottfredson, S. 1987, p. 36)
>
> [T]he solution to improved violence prediction is the same as for the improvement of clinical predictions in general—the use of actuarial methods. (Epperson, et al., 1995)

Based on a meta-analysis (52 studies of 16,191 persons) of predictors of general and violent recidivism, Bonta & Hanson (1998) found criminal history was the best predictor and clinical factors were the worst. (Also see Andrews & Bonta, 1998; Dawes, Faust, & Meehl, 1993; Gottfredson & Gottfredson, 1994; Grove & Meehl, 1996; Grubin, 1999; Hanson, 1998; Hanson & Bussiere, 1998; Hanson & Harris, 2000; Hanson & Thornton, 1999; Howe, 1994; Janus & Meehl; 1997; Milner & Campbell, 1995; Monahan, 1984, 1992, 1995, 1996; Webster, 1994; Quinsey & Maguire, 1986; Quinsey, et al., 1998).

In a more recent meta-analysis of 136 studies—i.e., *all* of the studies that the authors were able to find that compared clinical versus mechanical prediction and met minimum standards that would enable a comparison to be made, Grove, Zald, Lebow, Snitz, & Nelson (2000) concluded that "mechanical predictions of human behaviors are equal to or superior to clinical prediction for a wide range of circumstances" (p. 19). Note that:

> The only design variable [of the studies] that substantially influenced the relative efficacy of the mechanical- and clinical-prediction methods was whether the clinicians had access to a clinical interview. *Alas, clinical predictions were*

> *outperformed by a substantially greater margin when such data was available to the clinician.* (p. 25, emphasis added)

The 136 studies examined attempts to predict phenomena as diverse as the diagnosis of small bowel disease, business startup success, homosexuality, career satisfaction, and performance in medical school. The authors categorized the studies as falling into one of the following six categories: educational, financial, forensic, medical, clinical-personality, and other. Note that the single category that showed *the largest difference between clinical and mechanical prediction*, i.e., the *greatest* superiority of mechanical prediction, was in the forensic category. Of the 136 studies, there were ten studies that compared clinical and mechanical prediction in forensic settings and two looked at issues that might have forensic implications. These 12 studies were described as attempts to predict probation success, criminal behavior, parole success or failure, assaultive behavior, lie detection, juvenile delinquency, assault by psychiatric inpatients, juvenile criminal recidivism, and 4 studies of criminal recidivism. In 11 out of the 12 studies, mechanical prediction was superior to clinical judgment. In one study of probation success, clinical judgment yielded a 96% hit rate beating mechanical prediction, which yielded a 95% hit rate. Given a choice, the empirical data is unequivocally clear: (1) in forensic settings, you are almost certain to do better with mechanical prediction; (2) in the rare situation in which this is not so, you will do as well; and (3) any reliance on clinical interviews makes the clinical method even less accurate.

And there is evidence that when using non-actuarial methods—i.e., clinical, subjective judgment—that clinicians are no better than and often are *worse than lay persons*. For example, given enough information, laypersons making clinical predictions are as accurate as clinicians, and when there is a difference, lay people usually make better predictions.

> In the particular case of predicting violence, it is well documented that mental health professionals possess no special expertise in the prediction of violence, and that reliance on clinical judgements alone results in numerous inaccurate predictions of violent recidivism. (Rice and Harris, 1995)

Menzies, Webster, McMain, Staley, & Scaglione (1994) studied the accuracy of clinician and layperson predictions of dangerousness among Metropolitan Toronto Forensic Service patients using the Dangerous Behavior Rating Scale (DBRS). Three outcome measures were used: violent behaviour, criminal behaviour and general incidents. They found that clinicians were no better than laypersons at assessing risk; in fact, laypersons were better at using the DBRS than clinicians. (Also see Dawes, et al., 1989; Goldberg, 1959; Jackson, 2004; Menzies, Webster, & Sepejak, 1985; Monahan, 1981; Faust & Ziskin, 1988; Mossman, 1994).

Hanson (1998), a pioneer of and outspoken advocate of the statistical model, noted that guided clinical assessment cannot be ruled out because there were two (2) studies that did almost as well as the actuarial studies. In guided clinical assessment—in contrast to pure clinical method—only factors that have received empirical support are assessed. How the empirically validated factors are then weighted to reach a decision is left to clinical judgment. However, the larger of the two "guided clinical" studies Hanson cited was Epperson, et al.'s development study of the Minnesota Sex Offender Screening Tool (MnSOST). The MnSOST was actually an actuarial measure in which the evaluators reviewed the files in order to score the items in the actuarial tool. When they were done and had scored the MnSOST—i.e., after they had worked through each of the variables assessed in the actuarial tool and knew exactly what the actuarial measure predicted—they were asked to use clinical judgment to modify the actuarial prediction if they thought it should be adjusted. This was therefore an adjusted actuarial tool, not a guided clinical measure, using Hanson's own definitions.[3] The correlation between the actuarial MnSOST score and recidivism was .27. In accord with the view I am presenting, when the evaluators were allowed to use their judgment to adjust the actuarial prediction, accuracy fell to .20.

So, the larger of the two studies supporting the guided clinical method was actually an adjusted actuarial study that, once again, showed the superiority of actuarial over clinical input. Furthermore, the MnSOST has been replaced by the MnSOST-R. The details of the revision are quite telling in this regard:

> The second element of this revision was a change to empirical methods for item selection and scoring. The previous MnSOST produced a total score that was used in an actuarial manner, but the scoring of individual items was clinically based. In contrast, the MnSOST-Revised (MnSOST-R) utilized empirical methods for item selection and scoring. Given the strong support for the general superiority of systematically derived empirical risk assessments over intuitive or even trained clinical predictions, as summarized above, it was assumed that the predictive validity of the MnSOST-R would be significantly improved by using empirically based, rather than clinically based, item selection and scoring. (Epperson, et al., 1999)

This proved to be the case. So, the single major "guided clinical" assessment tool has been *improved by eliminating clinical judgment entirely*. And the accuracy of the only other study to show validity for guided clinical judgments was accidentally exaggerated.[4]

[3] When I brought this to Hanson's attention, he agreed:

> On reflection, I agree with your characterisation of the 1995 MnSOST study as an "adjusted actuarial" approach. The evaluators had scored the MnSOST, and they were allowed to form their own overall judgement. It was not clear to me on reading the original report whether the MnSOST scores were calculated by the raters or whether they only rated the items. Epperson's explanation seems reasonable and I will change my characterisation of the study accordingly. (Hanson, personal communication, emailed 5/21/01)

[4] A calculation error led Hanson to the conclusion that the accuracy of that study was .29 when in actuality it was .12 (the error accidentally multiplied the validity of that study by a factor of 6, from 1½% of the variance accounted for to 9%):

> I had originally calculated the accuracy of the Smith and Monastersky study as about .27, but I made a calculation error (pointed out to me by Grant Harris). The correlation is closer to .17 - I would have to go back and check the exact number, but it was less than .20. (Hanson, personal communication, via email, 5/14/01).
>
> I checked my notes on Smith & Monastersky. The original (erroneous) correlation was .29. It should be .12. This was a simple calculation error that was only pointed out to me after the 1998 article was published. *The actual correlation was pretty much what you would expect from clinical assessments.* (Hanson, personal communication, via email, 5/23/01, emphasis added).

One reason why psychologists predict poorly is the result of using factors that have been shown empirically to be *unrelated* to future events; clinical experts typically use their theory of what they *believe* (without corroborating empirical evidence) predicts future behavior. For example, experts frequently opine that *denial* of the offenses or details of the offenses indicates sexual dangerousness. Or that clinicians' evaluations of whether or not the offender shows *empathy* for his victims can be used to make accurate predictions. However, when one looks at all the studies that have evaluated denial and empathy prior to release and then looked at whether there was any difference between the recidivism rates of deniers (vs. admitters) or empathizers (vs. those that lack empathy), one finds no correlation.[5] The reality is that clinicians' evaluations of such things have no predictive utility whatsoever despite their clear beliefs to the contrary:

> [M]any of the frequently used risk assessment procedures have questionable validity . . . [T]he accuracy of [clinical] risk assessments has been unimpressive . . . The clinical prediction of sexual offender recidivism is no exception. Across 10 studies (N = 1,453) that examined the predictive accuracy of clinical judgments concerning sexual offender recidivism risk, the average correlation was only .10[6]

Taken with his conclusions that the MnSOST study was adjusted actuarial and not guided clinical (see discussion above), there are no findings in Hanson & Bussiere's 1998 meta-analysis (or at that time, anywhere else) that offered any support for guided clinical assessments (and, again, the states' experts typically use *unguided* clinical assessments that are even less accurate, though they now often call their method the "guided clinical method").

[5] Hanson, R. K. and Bussiere, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology*, 66, 348-362. (Used statistical techniques to combine all known studies on factors related to sex offender recidivism to determine what the empirical evidence tells us about what factors are predictive and how reliable they are. The 61 studies included 23,393 offenders. This overview is thus the "state of the art" summary of what we know about predicting sex offender recidivism.)

[6] "Correlations less than .10 would have little practical utility in most settings" (Hanson and Bussiere, 1998, p. 351). In actuality, Hanson and Bussiere found a correlation of .07 with clinical judgment when the two studies using "guided clinical judgment" were removed. Those studies had an average *r* = .23. However, the larger of the two "guided clinical" studies was Epperson, Kaul, and Huot's (1995) MnSOST development research. For technical reasons (because of oversampling of recidivists in their development sample), the correlations between measure and recidivism can be assumed to be inflated. Furthermore, Epperson (personal communication, May, 2001) noted that the "clinical judgment" was tightly controlled and "highly guided": The clinical judges were the scorers of the MnSOST and were forced to focus on the empirically validated factors, to carefully assess them, and to derive the actuarial score *prior* to using any clinical judgment. Because clinical judgment *lowered* accuracy (.27 without and .20 with clinical judgment), Epperson, et al. removed it when they developed the revised version, the MnSOST-R, a pure actuarial tool (Epperson, Kaul, and Hesselton, 1999) was not really a "guided clinical" assessment as Hanson (1998) defined the term. The MnSOST was actually an adjusted actuarial tool that was improved, in part, by *removing* all clinical judgments (eliminating the adjustment) when it was updated to become the MnSOST-R. When I brought this to Hanson's attention, he agreed with my classification of the studies.

While adjusting actuarial assessments may at times be necessary and can be reasonably defended in certain circumstances, there is no empirical evidence that this improves accuracy and considerable evidence that it decreases accuracy (Goldberg, 1959; Quinsey & Maguire, 1986; Dawes, Faust, and Meehl, 1989; Webster, Harris, Rice, Cormier, & Quinsey, 1994; Quinsey, Harris, Rice, & Cormier, 1998). For another example, consider the best predictor of non-sexual violence, the Violence Risk Appraisal Guide (VRAG). The VRAG was also intended to be an adjusted actuarial tool, when Webster, et al. (1994) created it. In less than five years, they recanted and eliminated all clinical adjustment to their measure (Quinsey, Harris, Rice, & Cormier, 1998). This will be discussed further. At this point, it is important to simply note that the type of clinical judgment typically used by State examiners has a correlation estimated to be .07 with recidivism (Hanson & Bussiere, 1998). This is so close to zero it can often be surpassed by tossing a coin, a fact that is consistent with the fact that lay people are usually as good as or better predictors than experts using clinical judgment.

> . . . Evaluators assessing the long-term risk for recidivism can be reasonably sure that a factor should not be considered when the average correlation is near 0 . . . Included among the unrelated factors are measures of general psychological adjustment . . . (average *r* = -.01) . . . Additionally, those offenders who denied their offense were at no higher risk for recidivism than other sexual offenders (average *r* =.02) . . . Many of the factors that clinicians intuitively believe to be related to sexual offense recidivism, such as denial of the offense and verbal statements of treatment motivation, have not been found to predict sexual offense recidivism[7]

A negative clinical presentation (e.g., low remorse, denial, low victim empathy) was unrelated to sexual recidivism.[8]

In summary, the clinical method for predicting human behavior is no longer taken seriously among

[7] Hanson, R. K. (1998). What do we know about sex offender risk assessment? *Psychology, Public Policy, and Law*, 4, 50-72.

[8] Hanson and Bussiere, 1998, p. 357.